基于线性规划的多类支持向量机算法*)

孙德山[1] 吴今培[2]

(辽宁师范大学数学学院 大连 116029)[1] (五邑大学智能技术与系统研究所 江门 529020)[2]

摘 要 多类支持向量机一般采用多个两类分类支持向量机来求解,这就需要解多个二次规划问题,从而导致算法的计算复杂性很高。根据一类分类思想,提出一种基于线性规划的多类分类算法及其分解形式,所给算法通过引入核函数能够独立地对每一类样本形成一个紧致的优化区域,从而达到分类的目的。对人工三螺旋线数据和几组实际数据库的识别实验表明,所给算法在保持良好的分类精度前提下,能有效地降低程序的运行时间。

关键词 线性规划,多类分类,一类分类,核函数

A Multi-class Support Vector Machine Algorithm Based on Linear Programming

SUN De-Shan[1] WU Jin-Pei[2]

(Institute of Math, Liaoning Normal University, Dalian 116029)[1]

(Institute of Intelligent Technology & System, Wuyi Uniersity, Jiangmen 529020)[2]

Abstract The multi-class support vector machine is commonly solved by decomposition to several binary support vector machines, which can bring complicated computation due to solving many quadratic programming problems. A multi-class classification algorithm and its decomposition form based on linear programming are proposed according to one-class classification idea in this paper, which can form a compact boundary about every single class sample by using kernel function and accordingly obtain the aim of classification. Simulations are conducted on artificial three spiral data and several real databases, which show that the proposed method can reduce the running time of program and guarantee good classification precision.

Keywords Linear programming, Multi-class classification, One-class classification, Kernel function

1 引言

在统计学习理论[1]基础上发展起来的支持向量机(Support Vector Machine, SVM)已经成为继神经网络之后机器学习的一个新的研究热点。支持向量机改变了传统的经验风险最小化原则,而是针对结构风险最小化原则提出的,因此具有很好的泛化能力。另外,支持向量机还通过引入正定核函数的方法,有效地克服了维数灾难及局部极小问题,在处理非线性问题时具有其它方法不可比拟的优越性。

由于支持向量机[2]最初是针对两类分类问题提出的,因此,存在一个如何将其推广到多类分类问题上。目前常用的方法有一对多方法、一对一方法、SVM决策树法[3]、有向无环图法[4,5]等,这些做法需要构造多个两类分类支持向量机,因此需要解多次二次规划问题,并且测试时需要对每两类都进行比较,导致算法计算复杂度很高。Weston 在 1998 年提出一种多类分类算法[6],它是直接在目标函数上进行改进,用一个优化式建立了多类分类支持向量机。但这种算法由于变量数目过多,因此只能在小型问题的求解中使用。

目前,针对异常值检测而提出的一类支持向量机方法[7]也采用了核方法,并且能很好地处理非线性问题。文[8]根据一类分类思想在二次规划的基础上提出了一种多类分类方法,该方法有效地克服了通过构造一系列两类分类支持向量机而带来的复杂计算问题,并且取得了满意的识别效果。为了进一步提高算法的运行速度,本文在文[7]的基础上,提出一种基于线性规划的多类分类方法。文章最后给出了几组对比实验来验证本文算法的有效性。

2 基于线性规划的多类分类算法

2.1 基于线性规划的一类支持向量机

给定一个正类样本点集$\{x_i, i=1,\cdots,l\}$,$x_i \in R^d$,用一个非线性映射ϕ将样本点映射到高维特征空间。一类支持向量机(1-SVM)的目的就是要在高维空间中找一个超平面[9],使之以尽可能大的距离ρ将尽可能多的样本从原点分离开,即估计一个函数$f_w(x)=\langle w,\phi(x)\rangle$,当一个样本$x$满足$f_w(x) \geqslant \rho$时,它被确定属于该类。为了获得$w$和$\rho$的值,采用下面的二次优化:

$$\min \frac{1}{2}\|w\|_2^2-\rho+C\sum_{i=1}^{l}\xi_i \tag{1}$$

约束为

$$\langle w,\phi(x_i)\rangle \geqslant \rho-\xi_i, \xi_i \geqslant 0, i=1,\cdots,l \tag{2}$$

其中,$\frac{1}{2}\|w\|_2^2$为规划项,参数C对误差项和规划项做出折中。

一类分类的另外一种途径是在高维空间中寻找包含尽可能多的样本的最小超球体[10]。当采用高斯核函数时,可以证明两种方法是等价的。另外,由于高斯核函数在实际应用中具有较好的性能,本文只以高斯核函数为例,它的定义如下:

$$k(x,y)=\exp\left(-\frac{\|x-y\|^2}{\sigma^2}\right) \tag{3}$$

若优化问题(1)中的规划项采用l_∞-范数,并根据文[11]中的定理 2.2 可以得到等价的优化问题:

*)基金项目:国家自然科学基金(70471074);广东省自然科学基金(032353)。孙德山 副教授,博士,主要研究方向:智能信息处理;吴今培 教授,博士生导师,主要研究方向:人工智能及电子技术。

$$\min -\rho+C\sum_{i=1}^{l}\xi_i \tag{4}$$

约束为

$$\langle w,\phi(x_i)\rangle\geqslant\rho-\xi_i,\xi_i\geqslant 0,i=1,\cdots,l \tag{5}$$

$$\|w\|_1=1 \tag{6}$$

采用核展开式$\sum_{j=1}^{l}a_jk(x_j,x_i)$代替优化问题(4)中的不等式约束项$\langle w,\phi(x_i)\rangle$，在不改变原优化问题的情况下，不妨设定$a_i\geqslant 0,i=1,\cdots,l$，于是得到下面的线性规划：

$$\min -\rho+C\sum_{i=1}^{l}\xi_i \tag{7}$$

约束为

$$\sum_{i=1}^{l}a_ik(x_i,x_j)\geqslant\rho-\xi_j,j=1,\cdots,l \tag{8}$$

$$\sum_{i=1}^{l}a_i=1 \tag{9}$$

$$a_i,\xi_i\geqslant 0,i=1,\cdots,l \tag{10}$$

解这个线性规划可以获得和的值，于是得到一个决策函数：

$$f(x)=\sum_{i=1}^{l}a_ik(x_i,x) \tag{11}$$

决策超平面$\sum_{i=1}^{l}a_ik(x_i,x)=\rho$映回到原空间后，就成为包含训练样本的紧致区域。对于区域内的任意样本x，满足$f(x)\geqslant\rho$，而对于区域外的任意样本y，将满足$f(y)<\rho$。

根据优化问题的意义，对于大部分训练样本将满足$f(x)\geqslant\rho$。参数C的意义就是控制满足条件$f(x)\geqslant\rho$的样本数量，参数C的值越小，区域外面的样本个数越多。图1是当参数$C=1$时的封闭曲线，此时所有样本点都在封闭曲线内。当参数$C=1/20$时，有一 部分样本跑到了优化区域的外面，如图2所示。核函数中的参数σ^2的取值越小，获得原空间中包含训练样本的区域越紧致，如图3(a)～(d)所示，这是进行分类的关键，同时也说明参数σ^2的大小将决定分类的精度，这在后面的实验中可以得到验证。

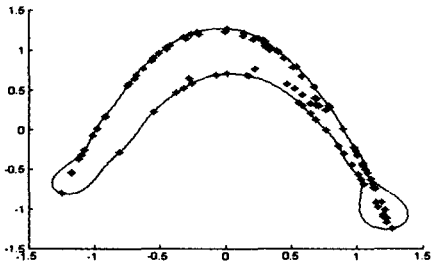


图1 $C=1,\sigma^2=0.2$时的封闭曲线

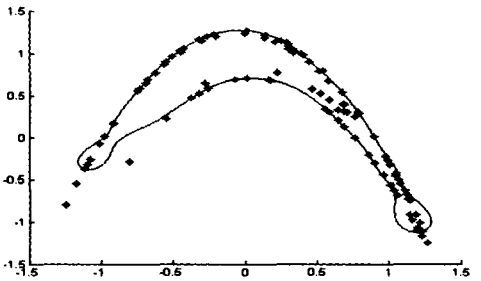
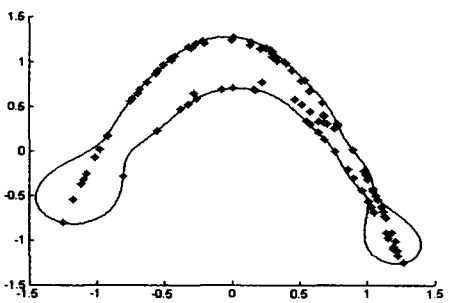
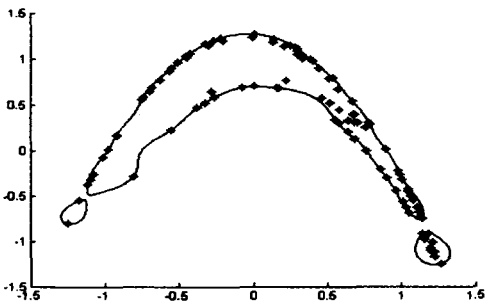


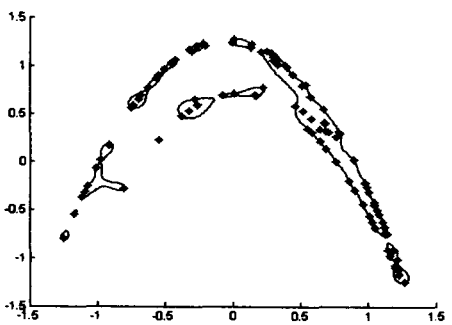
图2 $C=1/20,\sigma^2=0.2$时的封闭曲线



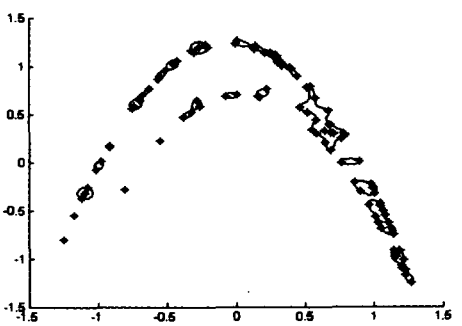
(a) $\sigma^2=0.3$



(b) $\sigma^2=0.1$



(c) $\sigma^2=0.03$



(d) $\sigma^2=0.01$

图3 参数σ的大小对边界区域的影响

2.2 多类分类算法

下面将线性规划下的一类分类方法推广到多类分类情况。采用的方法是对每一类样本实行一类分类运算，进而对每类样本都获得一个决策函数$f_i(x)$，然后将待识别样本输入每一个决策函数中，根据决策函数的最大值来判断该点所属的类，具体算法如下所述。

设训练样本为$\{(x_1,y_1),\cdots,(x_l,y_l)\}\subset R^n\times R$，其中，$n$为输入向量维数，$y_i\in\{1,2,\cdots,M\}$，$M$为类别数。将样本分成$M$类，各类分开写成，$\{(x_1^{(s)},y_1^{(s)}),\cdots,(x_{l_s}^{(s)},y_{l_s}^{(s)}),s=1,\cdots,M\}$其中，$\{(x_i^{(s)},y_i^{(s)}),i=1,\cdots,l_s\}$代表第$s$类训练样本，$l_1+\cdots+l_M=l$。根据一类分类思想，构造下面的线性规划：

$$\min -\sum_{s=1}^{M}\rho_s + C\sum_{s=1}^{M}\sum_{i=1}^{l_s}\xi_{si} \tag{12}$$

约束为

$$\sum_{j=1}^{l_s}\alpha_j^{(s)}k(x_j^{(s)},x_i^{(s)}) \geqslant \rho_s - \xi_{si}, s=1,\cdots,M, i=1,\cdots,l_s \tag{13}$$

$$\sum_{j=1}^{l_s}\alpha_j^{(s)}=1, s=1,\cdots,M \tag{14}$$

$$\alpha_i^{(s)},\xi_{si}\geqslant 0, s=1,\cdots,M, i=1,\cdots,l_s \tag{15}$$

解这个线性规划，可以获得 M 个决策函数 $f_s(x)=\sum_{j=1}^{l_s}\alpha_j^{(s)}k(x_j^{(s)},x), s=1,\cdots,M$。

给定待识别样本 z，计算 $\gamma_i=f_i(z), i=1,\cdots,M$，比较大小，找出最大的 γ_k，则 z 属于第 k 类。同时可定义该分类结果的信任度如下：

$$B_k=\begin{cases}1, \text{当 } \gamma_k\geqslant\rho_k\\ \dfrac{\gamma_k}{\rho_i}, \text{否则}\end{cases} \tag{16}$$

综上所述，该多类分类算法就是解一个线性规划问题，我们将该算法记为 LPMSVM。

2.3 分解算法

线性规划下的多类分类算法与标准支持向量机算法相比大大减少了计算量，但当数据量很大时，由于空间存储能力的限制，该算法在一般优化软件中仍然难以实现。根据线性规划(12)的特点，关于每一类样本的优化参数是独立的，因此可以将每一类样本分开优化来减少存储容量。另外，当每一类的样本数仍然很大时，还可以进一步采用分解方法来解决，具体步骤如下所述。

设整个训练样本为 M 类，分别将每一类的样本分解为一些较小规模的样本集合，设各类的分解数量分别为：$N_1, N_2, \cdots, N_M$，然后分别对各个小规模的数据样本实行一类分类优化操作，从而获得 N 个决策函数：

$$f_{ij}(x), i=1,\cdots,M, j=1,\cdots,N_i \tag{17}$$

其中，$N=N_1+\cdots+N_M$。

给定待识别样本 z，计算 $\gamma_i=(f_{i1}(z)+\cdots+f_{iN_i}(z))/N_i, i=1,\cdots,M$，比较大小，找出最大的 γ_k，则 z 属于第 k 类。

由于分解算法的优化过程是独立的，因此可以进行并行运算，将分解算法记为 LPDMSVM。

3 实例分析

本节采用三个实验来验证本文所给算法的有效性，第一个实验是人工的三螺旋线分类问题，另外两个实验是几组实际的分类数据，并同其它一些方法的实验结果进行了比较。

实验 1：人工三螺旋线数据

对于双螺旋线分类采用任何一个线性分类器对它只能有一半的分辨率。而对于三螺旋线的分类采用任何一个线性分类器对它只能有 33%的分辨率，它们经常被用作检验模式识别算法性能的试金石。下面以三螺旋线分类为例验证所给算法的分类效果，数据如图 4 所示，它是由下面三个式子产生：

(1) 第一类样本由下式产生：

$x1(t)=\exp(-0.2t)\cos(2t)$；$y1(t)=\exp(-0.2t)\sin(2t)$

(2) 第二类样本由下式产生：

$x2(t)=\exp(-0.2(t+1.5))\cos(2t)$；$y2(t)=\exp(-0.2(t+1.5))\sin(2t)$

(3) 第三类样本由下式产生：

$x3(t)=\exp(-0.2(t+2.5))\cos(2t)$；$y3(t)=\exp(-0.2(t+2.5))\sin(2t)$

其中，t 从 0 到 10，间隔为 0.02。每一类取 100 个样本，共 300 个样本作为训练集，另外取 300 个不同于训练样本的测试样本，参数取 $C=1/2$，核函数取高斯函数。

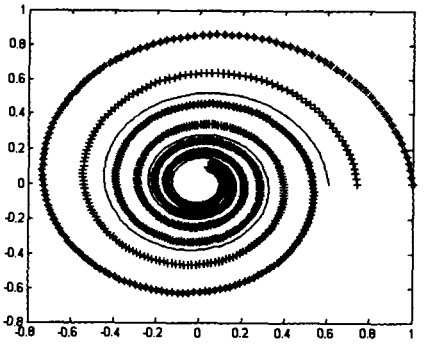


图 4 三螺旋线数据示意图

用 Matlab6.1 编程实现，采用不同的参数 σ，程序运行结果如表 1 所示。

表 1 核函数中的参数与识别率之间的关系

σ^2	0.1	0.05	0.03	0.01	0.005
识别率(%)	66	77	84	96	100

从表 1 中可以看出，随着 σ^2 的减小，识别精度逐渐增大，当 $\sigma^2=0.005$ 时，识别精度达到 100%，但并非 σ^2 的值越小越好，太小的值将导致无效的结果，该值可根据实际情况适当选取。

实验 2：实际数据

取公开的 UCI 数据：Wine 和 Glass，其中类别数分别为 3 类和 6 类，该数据可从网上下载获得[1]。

实验中取高斯核函数，其中，Wine 数据取 $\sigma^2=1.5$，Glass 数据取 $\sigma^2=0.008$，参数 C 都取 1。本文方法同文[5]中的方法比较结果如表 2 所示，其中，1-v-1(one-versus-one)代表用一对一方法构造的一系列 SVM 分类器，1-v-r(one-versus-rest)代表用一对多方法构造一系列 SVM 分类器，DAGSVM 是 John C. Platt 提出的利用有向无环图构造的多类 SVM 算法[82]，ACDMSVM 是文[5] 利用积极约束对 DAGSVM 方法的改进。

表 2 实验比较结果

数据集	1-v-1 测试正确率 运行时间(s)	DAGSVM 测试正确率 运行时间(s)	1-v-r 测试正确率 运行时间(s)	ACDMSVM 测试正确率 运行时间(s)	LPMSVM 测试正确率 运行时间(s)
Wine	95.2% 58.6	95.5% 58.8	95.4% 58.6	94.5% 55.6	97.78% 3.535
Glass	79.8% 114.3	82.3% 116.2	80.3% 103.3	78.3% 68.5	80.56% 3.044

从表 2 中可以看到，在取得较好分类精度前提下，本文方法的运行时间大大减少。

实验 3：数据 German、banana、breast-cancer

该数据来自于 www.first.gmd.de/～raetsch。German 有 700 个训练样本，300 个测试样本，20 维输入，输出维数是

[1] http://www.isc.uci.edu/~mlearn/MLRepository.html

1；banana 有 400 个训练样本，4900 个测试样本，2 维输入，输出维数是 1；Breast Cancer Data Set 有 200 个训练样本，77 个测试样本，输入维数是 9，输出维数是 1。这些数据都有 100 组训练样本和 100 组测试样本，这里任意选一组进行下面的实验。

首先采用文[7]中的二次规划下的方法进行实验，将该方法记为 QPMSVM。然后采用本文中的线性规划下的未分解算法（LPMSVM）和分解算法（LPDMSVM）进行对比实验。在分解算法中，将每类样本分成三组小样本集，然后对决策函数值采用加权平均，最后根据最大值判断新样本所属类别。

在 QPMSVM 方法中，取 $C=5,\sigma^2=0.3$。在本文的两种算法中采用相同的参数，其中，German 数据和 breast-cancer 数据中取 $C=5,\sigma^2=0.0001$，banana 数据中取 $C=5,\sigma^2=0.3$。三种算法的分类结果如表 3 所示。从表中可以看到，本文两种算法都保持了良好的分类精度，而且程序运行时间比二次规划下的算法运行时间大大缩短，尤其是分解算法运行速度更快。

表 3　实验结果

方法	German	banana	breast-cancer
QPMSVM	90% (125.5200)	87.29% (726.2550)	75.32% (16.4340)
LPMSVM	90% (78.2030)	88.86% (308.8140)	77.92%(7.2910)
LPDMSVM	90% (47.5080)	89.53% (248.7080)	77.92% (3.8850)

结论　本文根据一类分类思想，提出了一种基于线性规划的多类分类算法及其分解形式。对人工三螺旋线和实际数据的仿真实验结果表明该方法简单、运行速度快，而且仍然能保持良好的分类精度。当面对大规模数据分类问题时，可以采用分解算法来求解。另外，本文方法还定义了一个信任度函数，它可以在一定程度上对分类结果给出一个可信度评判，必要时可以根据信任度大小对分类结果进行适当调整。

参 考 文 献

1 Vapnik V N. Statistical Learning Theory [M]. New York, Wiley, 1998
2 Burges C J C. A Tutorial on Support Vector Machines for Pattern Recognition [R]. Knowledge Discovery and Data Mining, 1998, 2 (2): 121～167
3 Bennett K, Blue J. A support vector machine approach to decision trees [R]. Rensselaer Polytechnic Institute, Troy, NY: R. P. I Math Repot, 1997. 97～100
4 Platt J C, Cristianini N, Shawe-Taylor J. Large Margin DAGs for Multiclass Classification. In: Solla S A, Leen T K, Muller K R, eds. Advances in Neural Information Processing System 12 (NIPS 1999), Pittsburgh, PA, USA, Cambridge, MA, MIT Press, 1999. 547～553
5 李昆仑，黄厚宽，田盛丰. 一种基于有向无环图的多类 SVM 分类器. 模式识别与人工智能[J], 2003, 16(2): 164～168
6 Weston J, Watkins C. Multi-class Support Vector Machines [R]: [CSD-TR-98-04]. Royal Holloway University of London, 1998
7 孙德山，吴今培，肖健华. 一种新的多类分类算法. 模式识别与人工智能[J], 2004, 17(3): 357～361
8 Scholkopf B, Williamson R, Smola A, et al. Support Vector Method for Novelty Detection. http: // citeseer. nj. nec. com/ 400144. html
9 Ratsch G, Scholkopf B, Mika S, et al. SVM and Boosting: One Class. http: //citeseer. nj. nec. com/516656. html
10 Tax D. One-class classification: [PhD thesis]. Delft University of Technology, Netherlands, 2001
11 Mangasarian O L. Arbitrary-norm separating plane [J]. Operation Research Letters, 1999, 24(1): 15～23

（上接第 127 页）

所发现的负项对个数则随最小相关系数阈值的增大而减少。

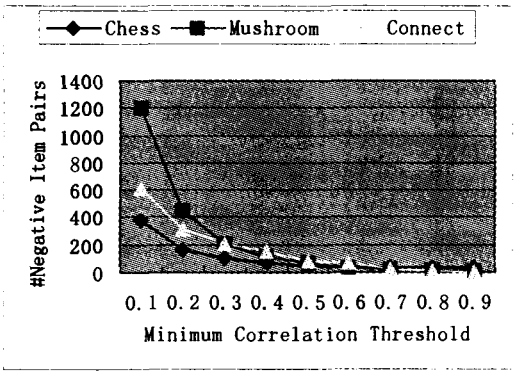


图 6　IBM 数据集负相关项对个数

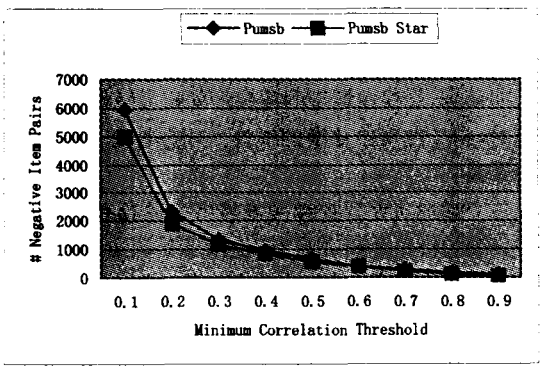


图 7　UCI 数据集负相关项对个数

下一阶段的研究将集中在算法的可扩展性和所发现的负规则的使用上。

参 考 文 献

1 Agrawal R, Srikant R. Fast algorithm for mining association rules in large databases. In: Proc. of 20th Int'l Conference on Very Large Databases (VLDB1994), 1994. 487～499
2 Han J, Pei J, Yin Y. Mining Frequent patterns without candidate generation. ACM SIGMOD, Dallas, Texas, 2000
3 Brin S, Motwani R, Silverstein C. Beyond Market Basket: Generalizing Association Rules to Correlation. In: Proc. 1997 ACM-SIGMOD Int'l Conf. Management of Data, 1998. 265～276
4 Savasere A, Omiecinski E, Navathe S. Mining for Strong Negative Associations in a Large Database of Customer Transactions. In: Proc. of the Fourteenth Intl. Conf. on Data Engineering (ICDE1998), 1998. 494 ～ 502
5 Ahmed K M, EI-Makky N M, Taha Y. A note on "Beyond Market Baskets: Generalizing Association Rules to Correlations". ACM SIGKDD Explorations, 2000, 1(2): 46～48
6 Wu X, Zhang C, Zhang S. Mining both positive and negative association rules. In: Proc. of 19th Intl. Conf. on Machine Learning (ICML2002), 2002. 658～665
7 Antonie M-L, Zaiane O R. An Associative Classifier based on Positive and Negative Rules. In: Proc. of the 9th ACM SIGMOD Workshop on Research Issues in Data Mining and Knowledge Discovery, 2004
8 Xiong H, Shekhar S, Tan P-N, et al. Exploiting A Support-based Upper Bound of Phi's Correlation Coefficient for Identifying Strongly Correlated Pairs. In: Proc. of the tenth ACM SIGKDD conference, 2004
9 http://www. almaden. ibm. com/software/quest/Resources/ index. shtml
10 http://www. ics. uci. edu/～mlearn/MLRepository. html